Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Защита информации и надёжность информационных систем

Студент: Скалкович С.Л.

ФИТ 3 курс 2 группа

Преподаватель: Нистюк О.А.

**Лабораторная работа №8**

**СЖАТИЕ/РАСПАКОВКА ДАННЫХ НА ОСНОВЕ СТАТИСТИЧЕСКИХ МЕТОДОВ**

Цель: приобретение практических навыков использования статистических методов Шеннона − Фано и Хаффмана (Shannon-Fano and Huffman coding) для сжатия/распаковки данных

Задачи:

1. Закрепить теоретические знания по алгебраическому описанию и использованию методов сжатия/распаковки (архивации/разархивации) данных на основе методов Шеннона − Фано и Хаффмана.

2. Разработать приложение для реализации методов Шеннона − Фано и Хаффмана.

1. **Теоретические сведения**

Статистические алгоритмы позволяют создавать более короткие коды для часто встречающихся и более длинные – для редко встречающихся символов алфавита или конкретного сообщения.

В первом случае метод считается статическим статистическим, во втором – динамическим статистическим: вероятностные свойства символов подсчитываются для конкретного сообщения или потока данных. Частота или вероятность появления того или иного символа алфавита в произвольном сообщении, лежащая в основе алгоритмов, дали название этим алгоритмам и соответствующим методам. Иногда эти методы называют также префиксными.

К примеру, если имеется некоторый код, который записывается как Х1 = А1А2, и другой код – Х2 = А1, то говорят, что Х2 является префиксом Х1. Или если Х1 = 1010, а Х2 = 10101100, то Х2 также является префиксом Х1. Таким образом, использование описываемых методов предусматривает создание кодовой таблицы (подобно кодам ASCII или base64). Формально процедура сжатия (прямое преобразование) состоит в подстановке соответствующего бинарного кода вместо символа исходного алфавита и наоборот – при обратном преобразовании

Код Шеннона – Фано не является оптимальным (обеспечивает минимальную избыточность) в общем смысле, хотя и дает оптимальные результаты при некоторых распределениях вероятностей. Для одного и того же распределения вероятностей можно построить, вообще говоря, несколько кодов Шеннона – Фано, и все они могут дать различные результаты.

Итак, необходимо выполнить следующие действия:

1) подсчитать вероятностные параметры символов алфавита А = {ai} (реализуется статическая версия алгоритма);

2) отсортировать – обычно в порядке убывания (невозрастания, т. е. могут иметь место повторяющиеся значения) вероятностей р(аi); р(аi) – вероятность появления в сжимаемом сообщении на произвольной позиции символа аi алфавита, т. е. создать таблицу символов алфавита, на основе которого генерируется сжимаемое сообщение;

3) каждому символу отсортированного множества поставить в соответствие бинарный код, для чего это множество (таблица) символов делится на две группы таким образом, чтобы каждая из групп имела приблизительно одинаковую суммарную частоту (вероятность). Очевидно, на первом шаге такая суммарная вероятность в каждой из групп должна быть максимально близка к 0,5. Первому из полученных подмножеств устанавливается первый символ бинарного кода: 0, второй − 1 (или наоборот). Для вычисления следующих битов кодов данная процедура повторяется рекурсивно для каждого из полученных на текущем шаге подмножеств, в котором содержится больше одного символа. Получим таблицу, в которой длина кодовых комбинаций меняется от минимального (lmin) до максимального (lmax) значений.

1. Практическая часть

1. Разработать авторское приложение в соответствии с целью лабораторной работы.

2. С помощью приложения выполнить прямое и обратное преобразования сообщения, состоящего из собственных имени и фамилии. Сжатие/распаковка данных на основе статистических методов 93 Можно использовать любой из известных методов сортировки символов массива. Метод кодировки (Шеннона − Фано, Хаффмана) использовать по указанию преподавателя.

При этом таблица отсортированных символов строится:

а) на основе данных, полученных в лабораторной работе № 2;

б) динамически, на основе анализа сжимаемого сообщения.

Для выполнения этого задания был разработан код на Python, в котором использовался метод Шеннона-Фано и Хаффмана. Были разработаны функции для прямого и обратного преобразования.

В результате его работы мы можем увидеть статистику заданного текста. Она отображена на рисунке 1.

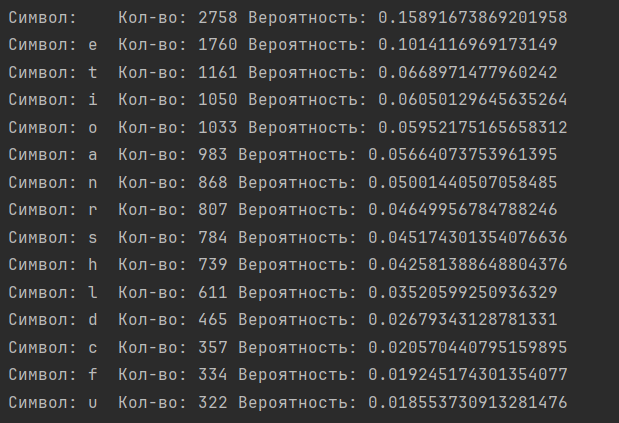


Рисунок 1 – Статистика количества и частоты символов в тексте

В зависимости от частоты встречи символов в тексте составляется таблица кодов, в которой каждому символу соответствует бинарная последовательность. Такая таблица представлена на рисунке 2.

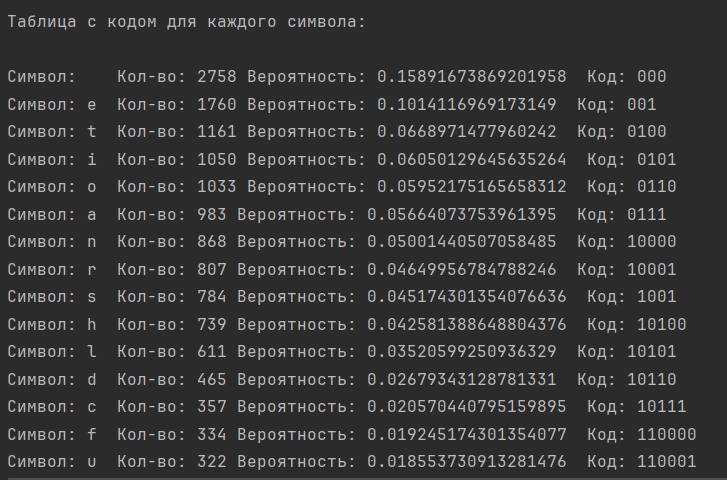


Рисунок 2 – Таблица кодов символов

В результате кодирования сообщения, его длина уменьшилась почти в два раза. Результат кодирования и декодирования представлен на рисунке 3.

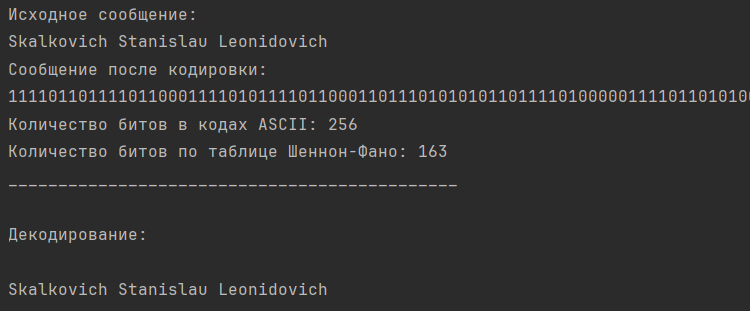


Рисунок 3 – Результат кодирования и декодирования

**Вывод:**

В ходе лабораторной работы были изучены два метода сжатия данных: Шеннона-Фано и Хаффмана. Метод Шеннона-Фано был реализован программно. Для оценки эффективности сжатия было проведено сравнение с использованием ASCII-кодирования путем подсчета символов в закодированном сообщении по методу Шеннона-Фано. Затем был посчитан коэффициент сжатия, который определяет, во сколько раз уменьшился общий объем данных. Результаты исследования показали, что применение метода Шеннона-Фано значительно снижает общий объем передаваемых данных.